CLAIMS

That which is claimed is:

1. A method of performing a domain-specific metasearch and obtaining search results therefrom, said method comprising the steps of:

providing a metasearch engine capable of accessing generic, web-based search engines and domain-relevant search engines;

receiving a query inputted by a user to the metasearch engine and searching for documents on a selected set of said generic, web-based search engines and domain-relevant search engines which are relevant to the query;

fetching raw data search results in the form of text documents from each member of the selected set;

displaying the raw data on a user interface;

supplying the raw data to a data mining module, wherein the data mining module forms clusters of related documents according to an unsupervised clustering procedure; and

displaying the clusters of related documents on the user interface.

2. The method of claim 1, wherein the unsupervised clustering procedure performed by the data mining module employs a group-average-linkage technique to determine relative distances between documents.

3. The method of claim 2, wherein the group-average-linkage technique employs the following algorithm for determining a proximity score that defines the relative distances between documents:

$$S_{ij} = 2 \times (1/2 - N(T_i,T_j)/(N(T_i) + N(T_j));$$

where T_i is a term in document i;

T_j is a term in document j;

$N(T_i,T_j)$ is the number of co-occurring terms that documents i and j have in common;

$N(T_i)$ is the number of terms found in document i; and

$N(T_j)$is the number of terms in document j.

4. The method of claim 1, wherein the data mining module, upon receiving the raw data, processes the raw data, independently of the unsupervised clustering procedure, and prepares a single list of all of the documents, after eliminating documents not reachable via the web.

5. The method of claim 4, wherein the data mining module assigns simple relevance scores to the documents prepared in the single list, based upon a frequency of terms from the query that appear within each of the documents.

6. The method of claim 5, wherein the documents are listed in the single list in an order ranging from a highest of the simple relevance scores to a lowest of the simple relevance scores.

7. The method of claim 1, further comprising the step of providing customized stop word lists to be used with regard to the generic, web-based search engines and domain-relevant search engines, wherein the data mining module references the stop word lists to strip stop words from documents associated with a respective generic, web-based engine or domain-relevant engine for which the particular stop word list being referred to has been customized, prior to determining the frequency of terms from the query that appear within each of the documents and computing a similarity score between results.

8. The method of claim 7, wherein the step of providing customized stop word lists comprises providing predefined customized stop word lists.

9. The method of claim 7, wherein the step of providing customized stop word lists comprises automatically generating stop word lists which are prepared and customized for each query.

10. The method of claim 5, further comprising displaying the single list on the user interface.

11. The method of claim 1, wherein the data mining module, upon receiving the raw data, processes the raw data, independently of the unsupervised clustering procedure, and categorizes the documents so that each document is assigned to one of a predefined number of categories.

12. The method of claim 11, further comprising providing a list of words for each of the categories wherein the words in each list are particular to the respective category, and wherein the data mining module compares the words in a particular list to a document to be characterized to determine whether the document is classified in that particular category.

13. The method of claim 12, wherein the step of providing a list of words for each of the categories comprises providing predefined lists.

14. The method of claim 12, wherein the step of providing a list of words for each of the categories comprises automatically generating the word lists which are prepared from a set of training documents.

15. The method of claim 14, wherein each word automatically selected for the generation of the word lists is identified based on a function computed from a frequency of occurrence of the word in the particular category for which it is selected, relative to a frequency of occurrence of the word in the other existing categories.

16. The method of claim 12, wherein the step of providing a list of words for each of the categories comprises automatically generating the word lists which are prepared by incremental training using previously selected lists of words and corresponding categories, as well as user feedback regarding the categorization of at least one of the documents.

17. The method of claim 11, wherein, upon completion of categorization of the documents, the documents are displayed in a categorized format to the user interface.

18. The method of claim 1, wherein the metasearch engine is further capable of accessing in-house, proprietary databases and any other informational databases that can be wrapped in a CGI-based web application server.

19. The method of claim 1, further comprising the steps of:

displaying a list of the generic search engines and domain-relevant search engines on the user interface which are available for searching; and

receiving a selection of all or part of the list from the user for directing the query thereto.

20. The method of claim 19, further comprising providing a context menu by which a user can select a group of search sites or engines by selecting a single context entry.

21. The method of claim 20, wherein the context menu includes at least one of the presets selected from the group consisting of a publications preset which selects more than one publications site, a sequences preset which selects more than one sequences site, a generic, web-based search engines preset which selects more than one generic, web-based search engine, a protein structure databases preset which selects more than one protein structure database, and a pathway information databases preset which selects more than one pathway information database.

22. The method of claim 1, wherein the documents consist of text-based data.

23. The method of claim 1, further comprising the steps of

storing at least one of the raw data and the clusters;

performing the steps of claim 1 to accomplish an additional search and data mining procedure;

storing at least one of the raw data and the clusters obtained from the additional search and data mining procedure;

receiving a sub-query inputted by a user to the metasearch engine and searching for documents from the data stored by the storing steps performed in regard to previous searches; which are relevant to the sub-query;

fetching raw data sub-query search results in the form of text documents from the stored data;

displaying the raw data sub-query search results on a user interface;

supplying the raw data sub-query search results to the data mining module, wherein the data mining module forms clusters of related documents according to an unsupervised clustering procedure; and

displaying the clusters of related documents resultant from the sub-query search on the user interface.

24. The method of claim 1, further comprising:

providing a browser including a relevance feedback mechanism;

analyzing the documents as they are browsed by a user on the user interface; and

generating a relevance weighting factor based upon observations resulting from the analyzing.

25. The method of claim 24, wherein the relevance weighting factor is applicable to a particular document having been browsed during the analyzing.

26. The method of claim 24, wherein the relevance weighting factor is applicable to a site or search engine from which a particular document having been browsed during the analyzing was fetched.

27. The method of claim 24, wherein the relevance weighting factor is applicable to a cluster in which a particular document having been browsed during the analyzing is grouped.

28. The method of claim 24, wherein the relevance weighting factor is applicable to a category in which a particular document having been browsed during the analyzing is categorized.

29. The method of claim 1, further comprising:

storing at least one of the raw data and the clusters;

performing the steps of claim 1 to accomplish an additional search and data mining procedure;

providing a browser including a relevance feedback mechanism;

analyzing the documents displayed from the additional search as they are browsed by a user on the user interface, wherein the analyzing includes comparing the documents being browsed with the stored data; and

generating a relevance weighting factor based upon observations resulting from the analyzing.

30. A method of performing a domain-specific metasearch and obtaining search results therefrom, the method comprising the steps of:

providing a metasearch engine capable of accessing generic, web-based search engines, publication sites, sequences sites, protein structure databases and pathway information databases;

receiving a query inputted by a user to the metasearch engine and searching for documents on a selected set of the generic, web-based search engines, publications sites, sequences sites, protein structure databases and pathway information databases which are relevant to the query;

fetching raw data search results in the form of text documents from each member of the selected set;

displaying the raw data search results on a user interface;

supplying the raw data to a data mining module, wherein the data mining module prepares a single list of all of the documents, after eliminating documents not reachable via the web, and assigns simple relevance scores to the documents prepared in the single list; forms clusters of related documents according to an unsupervised clustering procedure; and categorizes the documents so that each document is assigned to one of a predefined number of categories; and

displaying the documents in a format defined by the single list, in a format defined by the clusters, and in a format defined by the categories on the user interface so that a user can choose to browse the documents according to the list format, cluster format or categories format.

31. A method of performing a domain-specific metasearch and obtaining search results therefrom, said method comprising the steps of:

providing a metasearch engine capable of accessing generic, web-based search engines and domain-relevant search engines;

receiving a query inputted by a user to the metasearch engine and searching for documents on a selected set of said generic, web-based search engines and domain-relevant search engines which are relevant to the query;

fetching raw data search results in the form of text documents from each member of the selected set;

supplying the raw data to a data mining module, wherein the data mining module forms clusters of related documents according to an unsupervised clustering procedure, and wherein the data mining module categorizes the documents so that each document is assigned to one of a predefined number of categories; and

displaying the documents in a format defined by the clusters, and in a format defined by the categories on a user interface so that a user can choose to browse the documents according to the cluster format or the categories format.

32. The method of claim 31, further comprising:

storing at least one of the raw data and the clusters;

performing the steps of claim 31 to accomplish an additional search and data mining procedure;

providing a browser including a relevance feedback mechanism;

analyzing the documents displayed from the additional search as they are browsed by a user on the user interface, wherein the analyzing includes comparing the documents being browsed with the stored data; and

generating a relevance weighting factor based upon observations resulting from the analyzing.

33. A computer system for searching both general and domain-specific information resources simultaneously pursuant to a user query and for obtaining organized search results therefrom, the system comprising:

a metasearch engine capable of accessing a plurality of sites including generic, web-based search engines and domain-relevant search engines, for receiving documents from said plurality of sites in response to the user query;

means for selecting particular search engines from a plurality of generic, web-based search engines and domain-relevant search engines that are presented to a user;

means for displaying the received documents to the user;

means for assembling the received documents from the plurality of sites searched by the selected particular search engines into a single list;

means for assigning relevance ranks to the received documents in the single list and for organizing the documents in the single list according to said relevance ranks;

means for clustering the received documents into clusters according to an unsupervised clustering procedure;

and means for displaying said single list and said clusters to the user.

34. The computer system of claim 33, wherein said means for assigning relevance ranks assigns the relevance rank based upon a frequency of occurrence of query terms in each of the received documents.

35. The computer system of claim 33, further comprising:

means for providing customized stop word lists to be used with regard to said generic, web-based search engines and domain-relevant search engines, wherein said means for assigning relevance ranks references said stop word lists to strip stop words from documents associated with a respective engine for which the particular stop word list being referred to has been customized, prior to determining a frequency of terms that appear within each said document, and wherein said terms are used to determine proximity scores between said documents.

36. The computer system of claim 33, wherein said unsupervised clustering procedure performed by means for clustering employs a group-average-linkage technique to determine relative distances between documents.

37. The computer system of claim 36, wherein said group-average-linkage technique employs the following algorithm for determining a proximity score that defines said relative distances between documents:

$$S_{ij} = 2 \times (1/2 - N(T_i,T_j)/(N(T_i) + N(T_j));$$

where T_i is a term in document i;

T_j is a term in document j;

$N(T_i,T_j)$ is the number of co-occurring terms that documents i and j have in common;

$N(T_i)$is the number of terms found in document i; and

$N(T_j)$is the number of terms in document j.

38. The computer system of claim 33, further comprising:

means for categorizing the received documents, so that each document is assigned to one of a predefined number of categories; and

means for displaying said categories and said documents assigned thereto to the user.

39. The computer system of claim 38, further comprising means for storing a list of words for each of said categories wherein said words in each list are particular to the respective category, and wherein said means for categorizing compares the words in a particular list to a document to be characterized to determine whether the document is classified in that particular category.

40. The computer system of claim 38, further comprising means for providing a predefined list of words for each of the categories.

41. The computer system of claim 38, further comprising means for automatically generating a word list for each of the categories.

42. The computer system of claim 41, wherein said word lists are prepared from a set of training documents.

43. The computer system of claim 41, wherein each word automatically selected for the generation of the word lists is identified based on a function computed from a frequency of

occurrence of the word in the particular category for which it is selected, relative to a frequency of occurrence of the word in the other existing categories.

44. The method of claim 41, wherein said word lists are prepared by incremental training using previously selected lists of words and corresponding categories, as well as user feedback regarding the categorization of at least one of the documents contained in at least one of the categories.

45. The computer system of claim 33, further comprising:

means for storing said received documents; and

means for performing a sub-query inputted by a user to search for documents stored by said means for storing which are relevant to said sub-query;

means for fetching raw data sub-query search results from said means for storing in the form of text documents;

means for displaying said raw data sub-query search results to the user;

means for assembling the raw data sub-query search results into a single list;

means for assigning relevance ranks to the raw data sub-query search results and for organizing the results in the single list according to said relevance ranks;

means for clustering the received sub-query documents into clusters according to an unsupervised clustering procedure;

and means for displaying said sub-query documents to the user in the single list and clusters formats.

46. The computer system of claim 33, further comprising:

a browser including a relevance feedback mechanism, adapted to analyze the documents as they are browsed by a user on a user interface; and to generate a relevance weighting factor based upon observations resulting from the analysis.

47. A computer system for searching both general and domain-specific information resources simultaneously pursuant to a user query and for obtaining organized search results therefrom, the system comprising:

a metasearch engine capable of accessing a plurality of sites including generic, web-based search engines and domain-relevant search engines for receiving documents from said plurality of sites in response to the user query;

means for selecting particular search engines from a plurality of generic, web-based search engines and domain-relevant search engines that are presented to a user;

means for clustering the received documents into clusters according to an unsupervised clustering procedure;

means for categorizing the received documents, so that each document is assigned to one of a predefined number of categories; and

means for displaying said clusters, said categories and said documents assigned thereto to the user.

48. The computer system of claim 47, further comprising:

means for displaying the received documents to the user;

means for assembling the received documents from the plurality of sites into a single list;

means for assigning relevance ranks to the received documents in the single list and for organizing the documents in the single list according to said relevance ranks;

means for storing said received documents; and

means for performing a sub-query inputted by a user to search for documents stored by said means for storing which are relevant to said sub-query;

means for fetching raw data sub-query search results from said means for storing in the form of text documents;

means for displaying said raw data sub-query search results to the user;

means for assembling the raw data sub-query search results into a single list;

means for assigning relevance ranks to the raw data sub-query search results and for organizing the results in the single list according to said relevance ranks;

means for clustering the received sub-query documents into clusters according to an unsupervised clustering procedure;

means for categorizing the received sub-query documents, so that each document is assigned to one of a predefined number of categories; and

means for displaying said sub-query documents to the user in the single list, categories and clusters formats.

49. The computer system of claim 47, further comprising:

a browser including a relevance feedback mechanism, adapted to analyze the documents as they are browsed by a user on a user interface; and to generate a relevance weighting factor based upon observations resulting from the analysis.

50. A computer readable medium carrying one or more sequences of instructions from a user of a computer system for searching both general and domain-specific information resources simultaneously to obtain organized search results therefrom, wherein execution of the one or more sequences of instructions by one or more processors causes the one or more processors to perform the steps of:

receiving a query inputted by the user and receiving instructions as to which databases to access;

accessing selected sites using generic, web-based search engines and domain-relevant search engines, based upon said instructions received from the user, and searching for documents on the selected sites, which are relevant to said query;

fetching raw data search results in the form of text documents from each of the selected sites;

displaying said raw data on a user interface;

forming clusters of related documents form said raw data, according to an unsupervised clustering procedure; and

displaying said clusters of related documents on the user interface.

51. The computer readable medium of claim 50, wherein the following further steps are performed:

preparing a single list of all of said documents, independently of said forming clusters, after eliminating documents not reachable via the web; and

assigning simple relevance scores to said documents prepared in said single list, based upon a frequency of terms from said query that appear within each said document.

52. The computer readable medium of claim 51, wherein the following further step is performed:

providing customized stop word lists to be used with regard to said generic, web-based search engines, publication sites and sequences sites, and referencing said stop word lists to strip stop words from documents associated with a respective engine, publication site or sequence site for which the particular stop word list being referred to has been customized, prior to determining said frequency of terms that appear within each said document and using the terms to compute proximity scores between the documents for clustering the documents.

53. The computer readable medium of claim 50, wherein the following further steps are performed:

processing said raw data, independently of said unsupervised clustering procedure, and categorizing said documents so that each document is assigned to one of a predefined number of categories.

54. The computer readable medium of claim 50, wherein the following further steps are performed:

providing a browser including a relevance feedback mechanism;

analyzing the documents as they are browsed by the user; and

generating a relevance weighting factor based upon observations resulting from said analyzing.